

Galaxy @ CRUK-CI

<http://galaxy.cruk.cam.ac.uk/>

Anne Pajon, Bioinformatics Core

Reproducible Research . User Experiences

Friday 27 November 2015

Definition

Reproducibility is the ability that an analysis described in sufficient detail **can** be precisely reproduced.
(by another person, in another environment)

Core tasks

1. Capture the **precise description** of the analysis
2. Assemble all the **necessary data** and **software dependencies** needed by the described analysis
3. Combine the above to verify the analysis

Key problems

Missing software, version, parameters (even data)

- **Tools**: inaccessible, hard to record details
- **Datasets**: not all available, difficult to access
- **Publication**: results, data, methods separate

50 papers citing BWA randomly selected from 378 published in 2011
Nature Reviews Genetics 13, 667-672 (September 2012) | doi:10.1038/nrg3305

- **31** provide **no** version and **no** settings
- **8** list versions, **4** list settings, **7** list versions and settings
- **26** do not provide access to data

Galaxy web-platform for bioinformatics analysis

- Integrate existing tools into a uniform framework
- **Accessible**
 - Users without programming experience can easily run tools
- **Reproducible**
 - History system tracks multisteps analysis
 - Any user can repeat and understand a complete computational analysis

The screenshot displays the Galaxy web-platform interface for the CRUK Cambridge Institute Galaxy server. The top navigation bar includes links for Analyze Data, Workflow, Shared Data, Lab, Visualization, Admin, Help, and User. The left sidebar contains a 'Tools' menu with categories like Get Data, Send Data, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Operate on Genomic Intervals, Statistics, Graph/Display Data, Unix Tools, FASTA manipulation, Multiple regression, Multivariate Analysis, NGS: SAM Tools, NGS: CHIP-seq, NGS: RNA-seq, NGS: Picard, NGS: BAM Tools, NGS: QC and manipulation, NGS: Mapping, NGS: Visualization, BIOINFORMATICS CORE, DNA Motif Tools, Other Tools, Plotting Tools, Heatmap Tools, Diagnostic Tests, Survival Analysis, Testing Tools, and Workflows. The main content area shows a message about a new authentication system and an update to the Galaxy server. The right sidebar displays a 'History' list with datasets like '5: RP-Taylor2010 on data.1' and '3: RP-Loi2007 on data.1', each with details on format, size, and database.

Galaxy / CRUK-CI

Analyze Data Workflow Shared Data Lab Visualization Admin Help User

Tools

search tools

Get Data
Send Data
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Operate on Genomic Intervals
Statistics
Graph/Display Data
Unix Tools
FASTA manipulation
Multiple regression
Multivariate Analysis
NGS: SAM Tools
NGS: CHIP-seq
NGS: RNA-seq
NGS: Picard
NGS: BAM Tools
NGS: QC and manipulation
NGS: Mapping
NGS: Visualization
BIOINFORMATICS CORE
DNA Motif Tools
Other Tools
Plotting Tools
Heatmap Tools
Diagnostic Tests
Survival Analysis
Testing Tools
Workflows
All workflows

CRUK Cambridge Institute Galaxy server

This is our local Galaxy server maintained by the Bioinformatics Core.

UPDATE - 27 October 2015

New Authentication System - Our Galaxy production server is now using the CRUK-CI authentication system. It means that from now on you have to use your cruk.cam.ac.uk email address and password to log in into Galaxy. Our Galaxy server is currently running the latest Galaxy distribution released July 2015 (v15.07). An FTP server has been configured to allow files larger than 2GB to be uploaded in Galaxy. Please see below in the 'To Start' section for more details.

To Start

Click on the *Get Data* link to the left and import some data.

For datasets larger than 2G, you will need to upload your data onto the FTP galaxy server first. To get started, you'll need to have registered a regular Galaxy account. Once registered, you can initiate an FTP connection in your preferred FTP client to galaxy.cruk.cam.ac.uk using your registered email address and password for the login details. Files uploaded to the FTP server won't automatically be imported to Galaxy - rather, you will be presented with a list of the contents of your FTP directory on the standard 'Upload File' tool interface. Files not imported within 3 days will be cleaned up from the FTP site.

Tools

The *Tools* menu allow you to load data into your Galaxy workspace and run a variety of analysis tools.

History

The *History* list contains data you import into Galaxy and the results of analysis tools you run. There you can delete data or edit their attributes, download the complete data using the "display" link, or use the "peek" feature to display the first few lines.

Help

If this is your first time on Galaxy and you would like a little direction to get started, please check out the [screencasts](#) and [documentation](#). We do also run from time to time an 'Introduction to Galaxy' course.

For any other questions, please contact [Anne Pajon](#) from the Bioinformatics Core. Thanks.

Galaxy is an open source, web-based platform for data intensive biomedical research. The Galaxy team is a part of the Center for Comparative Genomics at Penn State, and the Department of Biology at Johns Hopkins University. The Galaxy Project is supported in part by NSF, NHGRI, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Johns Hopkins University.

History

search datasets

Recursive partitioning
5 shown

1.0 MB

5: RP-Taylor2010 on data.1
1022 bytes
format: **html**, database: ?
HTML file

4: RP-Taylor2010 on data.1
55 lines
format: **tabular**, database: ?
HTML file

1	2	3	4
Gene	Accession	CutOff1	CutOff2
MUC5B	NM_002458	NA	NA
RAO51AP1	NM_006479	5.66	NA
PLK4	NM_014264	NA	NA
RAT1	NM_030665	NA	NA
AQP1	NM_198098	NA	NA

3: RP-Loi2007 on data.1
3.3 KB
format: **html**, database: ?
HTML file

2: RP-Loi2007 on data.1
74 lines
format: **tabular**, database: ?
HTML file

1	2	3	4
---	---	---	---

Galaxy tracks every steps of every analysis

- Exact parameters of a step can always be inspected, and easily rerun

RP-Taylor2010 Recursive partitioning on prostate cancer dataset (Galaxy Tool Version 1.0.0) Options

Source file
1: survival_genes.tabular
File containing a list of gene symbols, must be tabular format

Column
Column: 2
Column containing gene symbols

Surgical margin status
Negative
Filter expression data on surgical margin status

Groups
Primary
Filter on group either Primary or Primary + metastatic (all)

Gleason Grade
All
Filter on Gleason grade, 5-6 (low), 7-9 (high) or 5-9 (all)

Pathological T stage
All
Filter on pathological t-stage, T2, T3/T4 or all

Execute

4: RP-Taylor2010 on data 1

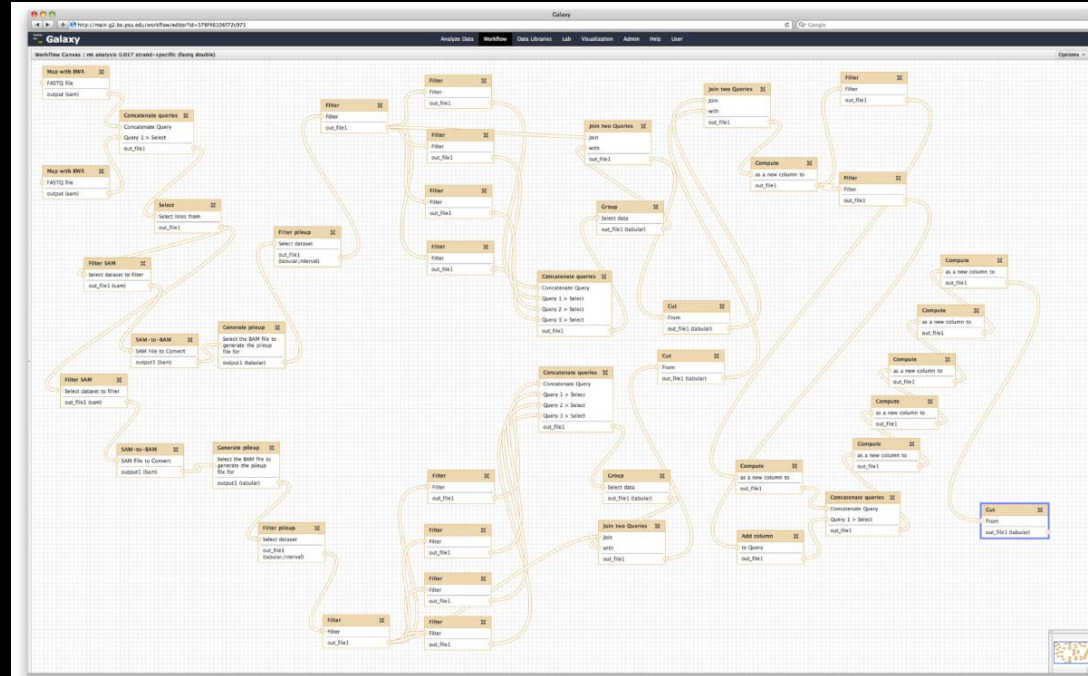
55 lines
format: **tabular**, database: ?

Run this job again 3 4

Gene	Accession	CutOff1	CutOff2
MUC5B	NM_002458	NA	NA
RAD51AP1	NM_006479	5.66	NA
PLK4	NM_014264	NA	NA
RAI1	NM_030665	NA	NA
AQP1	NM_198098	NA	NA

Galaxy workflow system

- **Workflows** can be constructed from scratch or extracted from existing analysis histories
- Facilitate reuse, as well as providing precise **reproducibility** of a complex analysis



Workflow for finding heteroplasmic sites from Illumina data.
This workflow can be accessed, used, and edited at <http://usegalaxy.org/heteroplasmy>.
Goto et al. *Genome Biology* 2011 12:R59 | doi:10.1186/gb-2011-12-6-r59


Galaxy data and tool sharing

Galaxy items - histories, workflows, visualizations, and pages - can be shared with other people

Galaxy tools - can be published and shared using the tool shed

<https://toolshed.g2.bx.psu.edu/>

Repository revision

7 (2015-10-09)  repository tip

Select a revision to inspect and download versions of Galaxy utilities from this repository.

Repository tophat2

Name: [tophat2](#)


Owner: [devteam](#)


Synopsis: Tophat - fast splice junction mapper for RNA-Seq reads

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie(2), and then analyzes the mapping results to identify splice junctions between exons.

Content homepage: <http://ccb.jhu.edu/software/tophat/index.shtml>

Development repository: <https://github.com/galaxyproject/tools-devteam/tree/master/tools/tophat2>

Link to this repository: <https://toolshed.g2.bx.psu.edu/view/devteam/tophat2/4eb3c3beb9c7> 

Clone this repository: hg clone <https://toolshed.g2.bx.psu.edu/repos/devteam/tophat2> 

Type: unrestricted

Revision: [7:4eb3c3beb9c7](#)

This revision can be installed: True

Times cloned / installed: 2125

Dependencies of this repository

Repository dependencies - installation of these additional repositories is required

Repository package [bowtie 2.2.5](#) revision [30bd7eaeddbf](#) owned by [iuc](#)

Repository package [tophat 2.0.14](#) revision [b13271391f95](#) owned by [iuc](#)

Tool dependencies - repository tools require handling of these dependencies

Name	Version	Type
bowtie2	2.2.5	package
tophat	2.0.14	package

Galaxy is an open-source, web-based, data integration and analysis platform for life science research.

Galaxy enables bench scientists to create, share, and publish sophisticated, **reproducible** bioinformatic analyses without requiring researchers to learn command line interfaces, or Unix system management skills.

Galaxy can be accessed through the project's public server, or on one of the over 60 publicly accessible Galaxy servers. Galaxy can also be installed locally, and on cloud infrastructures.

<https://usegalaxy.org/>

<http://galaxy.cruk.cam.ac.uk/>

<http://galaxycam.github.io/>